第 46 卷 第 2 期 计 算 机 工 程 2020 年 2 月
Vol. 46 No. 2 Computer Engineering February 2020

·热点与综述· 文章编号：1000-3428(2020)02-0011-10 文献标志码：A 中图分类号：TP301

网络流量异常检测中的维数约简研究

陈良臣[1,2,3]，高 曙[1]，刘宝旭[2]，陶明峰[4]

（1. 武汉理工大学 计算机科学与技术学院，武汉 430063；2. 中国科学院信息工程研究所，北京 100049；
3. 中国劳动关系学院 应用技术学院，北京 100048；4. 国网山东省电力公司 淄博供电公司，山东 淄博 255000）

摘 要：对包含大流量数据的高维度网络进行异常检测，必须加入维数约简处理以减轻系统在传输和存储方面的压力。介绍高速网络环境下网络流量异常检测过程以及维数约简方式，阐述流量数据常用特征和维数约简技术研究的最新进展。针对网络流量特征选择和流量特征提取 2 种特征降维方式，对现有算法进行归纳分类，分别描述算法原理及优缺点。此外，给出维数约简常用的数据集和评价指标，分析网络流量异常检测中维数约简技术研究面临的挑战，并对未来发展方向进行展望。

关键词：网络异常检测；流量维数约简；流量特征提取；流量特征选择；网络空间安全

开放科学（资源服务）标志码（OSID）：

中文引用格式：陈良臣，高曙，刘宝旭，等. 网络流量异常检测中的维数约简研究[J]. 计算机工程，2020，46(2)：11-20.
英文引用格式：CHEN Liangchen，GAO Shu，LIU Baoxu，et al. Research on dimensionality reduction in network traffic anomaly detection[J]. Computer Engineering，2020，46(2)：11-20.

Research on Dimensionality Reduction in Network Traffic Anomaly Detection

CHEN Liangchen[1,2,3]，GAO Shu[1]，LIU Baoxu[2]，TAO Mingfeng[4]

（1. School of Computer Science and Technology，Wuhan University of Technology，Wuhan 430063，China；
2. Institute of Information Engineering，Chinese Academy of Sciences，Beijing 100049，China；
3. School of Application Technology，China University of Labor Relations，Beijing 100048，China；
4. Zibo Power Supply Company，State Grid Shandong Electric Power Company，Zibo，Shandong 255000，China）

【Abstract】To implement anomaly detection for a high dimensional network with mass flow data，data dimensionality should be reduced to relieve transmission and storage burdens from the system. This paper introduces network traffic anomaly detection process and dimensionality reduction ways in hig-speed network environment. Then it summarizes common features of feature in network traffic anomaly detection and latest research developments of dimensionality reduction for traffic data. Aiming at two kinds of feature dimensionality reduction ways，network traffic feature selection and network traffic feature extraction，this paper lists and classifies frequently used algorithms and describes the principles，advantages and disadvantages respectively. On this basis，this paper analyzes existing datasets and evaluation indexes used in research of dimensionality reduction. Finally，this paper discusses development directions and challenges of dimensionality reduction technologies in network traffic anomaly detection.

【Key words】network anomaly detection；traffic dimensionality reduction；traffic feature extraction；traffic feature selection；cyberspace security

DOI：10.19678/j.issn.1000-3428.0056532

基金项目：国家自然科学基金（61802404，61602470）；国家信息安全专项（发改办高技[2015]289 号）；中国科学院战略性先导 C 类课题（XDC020400100）；中国劳动关系学院科研项目（20XYJS003，20ZYJS017）；北京市科委重点研究项目（D181100000618003）；中国科学院网络测评技术重点实验室基金；网络安全防护技术北京市重点实验室基金。

作者简介：陈良臣（1982—），男，讲师、博士研究生，主研方向为大数据、网络攻防技术、安全态势感知；高 曙，教授、博士、博士生导师；刘宝旭（通信作者），研究员、博士、博士生导师；陶明峰，教授级高级工程师。

收稿日期：2019-11-07 **修回日期：**2019-12-12 **E-mail：**liubaoxu@iie.ac.cn

0 概述

随着互联网技术的快速发展以及世界各国对网络信息化进程的加速推进,网络通信已渗透到各个领域,而互联网上的攻击手段也更隐蔽和智能,传统补丁式的网络安全解决方案无法完全解决日益暴露的安全问题[1]。针对网络流量的异常检测与监控已成为目前安全工具研究的主要方向。

在高速网络环境中,网络异常检测过程需要获取、处理和传输的大量网络流量数据,可能由大量特征来描述,通常这些特征中含有许多无关特征和冗余特征,会提高异常检测模型的复杂度,且各特征之间的相互干扰会导致检测性能急剧下降。因此,在对海量高维网络流量数据进行异常检测建模之前,需要对数据进行特征降维约简处理。攻击数据集的特征质量直接决定入侵检测系统的检测效率和稳定性,因此,分析网络流量以确定有助于识别攻击的维数约简方法至关重要。

针对基于网络流量的网络入侵异常检测模型,很多学者从网络流量特征选择和网络流量特征提取 2 个方面对维数约简问题进行研究。本文总结网络流量异常检测中流量数据常用特征和流量数据维数约简研究的最新进展,对网络流量异常检测中的网络流量特征选择方法和网络流量特征提取方法进行归纳分类,并列举常用算法、数据集和评价指标。在此基础上,阐述网络流量异常检测中维数约简技术研究面临的挑战,同时对未来发展方向进行展望。

1 网络流量异常检测与维数约简

网络流量指的是单位时间内网络上传输的信息量,即 2 个终端之间拥有相同通信五元组信息(源 IP 地址、源端口、目的 IP 地址、目的端口和传输层协议)的连续数据包[2]。在基于网络流量的异常检测过程中,需要对原始网络流量数据进行降维,从而有效提高异常检测算法的泛化能力[3]。

1.1 网络流量异常检测

入侵检测技术可分为误用检测和异常检测,其中异常检测基于与正常活动的显着偏差发现入侵[4]。网络流量异常检测就是分析从网络中采集的各种数据,挖掘结构中复杂和潜在的关系,从而推断出当前网络的安全状况,发现不可预见的攻击[5],其中主要包括两方面:1)提取网络流量数据中的关键信息作为异常检测的数据源;2)提取关键信息中的异常行为进行检测与识别[6]。通用的异常检测方法往往并不适用于网络流量。基于特征或行为、基于数理统计和基于流挖掘的网络流量异常检测方法已成为网络流量异常检测的主流和趋势。

网络流量异常检测过程如图 1 所示,可将其分为 5 个步骤,即网络流量数据获取、流量数据抽样、流量维数约简、异常检测建模以及异常检测结果与评估。

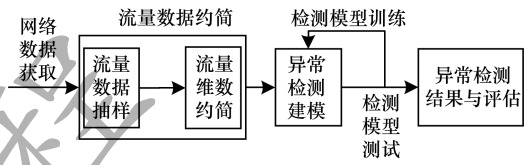

图 1 高速网络环境下的网络流量异常检测过程

Fig. 1 Network traffic anomaly detection process in a high-speed network environment

1.2 网络流量维数约简

维数约简又称为特征降维,网络流量维数约简一般包括网络流量特征选择和网络流量特征提取 2 种方式,两者都是为了从原始网络流量特征中找出最有效的特征[7],针对高维灾难都可以达到降维的目的,但是两者有所不同。网络流量特征选择是依据一定的规则从已有的网络流量特征中选取出部分特征来表示原始网络流量数据,如图 2(a)所示。网络流量特征选择保留了训练样本的原始物理意义,但是当网络流量数据间相似性很强时,检测冗余信息对计算要求非常高。网络流量特征提取则是按照一定的规则将原始网络流量特征空间变换成一个维数更小的空间,是使用数学方法对某些特征进行融合产生了新的特征,新的特征只具有数学含义,难以找到其现实意义,如图 2(b)所示。网络流量特征提取是在网络流量特征选择的基础上对网络流量数据集做进一步简化,去除剩余特征的冗余值[8-9]。

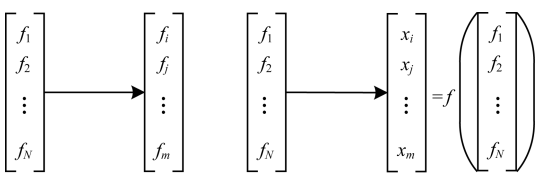

图 2 网络流量特征降维的 2 种方式

Fig. 2 Two ways to reduce the dimensionality of network traffic features

网络流量维数约简可以使网络流量数据集更容易使用,减少数据存储并降低算法的计算开销,同时

提高网络异常检测性能。为生成可靠的 IDS 模型,维数约简被认为是提高网络异常检测运算效率和发现数据模式的一项重要任务。

2　网络流量维数约简技术研究进展

维数约简算法中的“降维”,指的是降低特征矩阵中特征的数量。本节主要介绍网络流量异常检测中用到的特征归类研究和维数约简技术研究进展。

2.1　网络流量特征研究

网络流量异常检测中用到的网络流量特征大致可分为 3 类,即基于报文头部、基于网络流和基于连接图的网络流量特征[10],如图 3 所示,其中,基于报文头部的网络流量特征一般包含 IP 地址、端口地址等;基于网络流的网络流量特征主要是使用与网络流量相关的统计数据作为特征,即使用网络流的统计特征来表示网络流量,如包长、包到达间隔等,可进一步分为单流特征和多流特征;基于连接图的网络流量特征是图特征与网络流量特征相结合的网络流量特征。

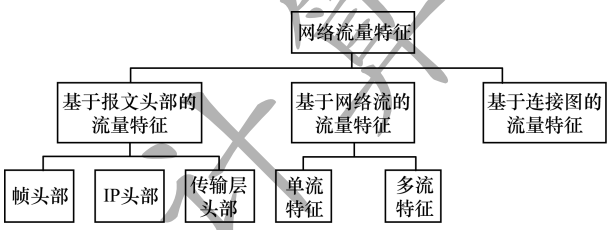

图 3　网络流量异常检测特征分类

Fig. 3　Classification of features used in network traffic anomaly detection features

网络流可分为单向流和双向流,网络流量特征也可分为单流特征和双流特征。单流特征即单个流的特征,只使用组成该网络流的所有报文集合的统计特征作为该网络流量的特征,通常包括包到达时间、报文大小、报文大小的均值/方差、网络流所包含的数据报文数量等。多流特征是针对具有某些相同特性的多条网络流量共同形成的一些统计特征,可在单流特征基础上表示出更多流量相关的信息。在网络流量异常检测过程中提取多流特征,一般先选择一个提取对象,如将主机地址作为对象的网络流量,或将网络段作为提取对象的网络流量等[10]。

2.2　网络流量维数约简技术研究

网络流量异常检测中的维数约简技术研究分类如图 4 所示。

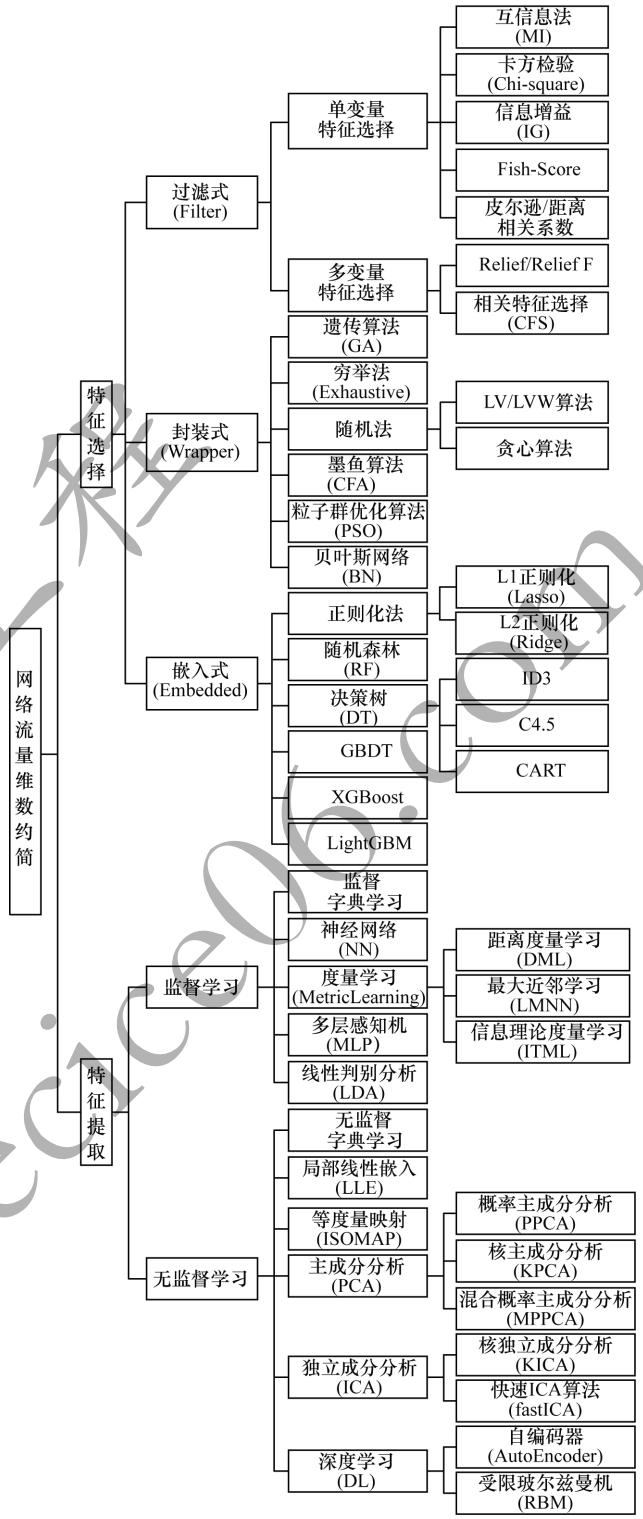

图 4　网络流量维数约简技术分类

Fig. 4　Classification of network traffic dimensionality reduction technologies

网络流量维数约简方法可分为网络流量特征选择方法和网络流量特征提取方法。网络流量特征提取维数约简方法可分为监督学习(Supervised)特征提取方法和无监督学习(Unsupervised)特征提取方

法。网络流量特征选择维数约简方法可分为嵌入式(Embedded)特征选择、过滤式(Filter)特征选择和封装式(Wrapper) 特征选择 3 种[11]。

3 网络流量特征提取方法

网络流量特征提取是通过功能映射,从原始网络流量特征集中提取一组新特征,该方法能够通过转换获取最小的新特征集[12]。

3.1 网络流量特征提取方法分类

网络流量特征提取方法包括有监督特征学习方法和无监督特征学习方法,其中有监督学习方法包括监督字典学习、神经网络、多层感知机、线性判别分析等,无监督学习方法包括无监督字典学习、局部线性嵌入、等度量映射、主成分分析(Principal Component Analysis,PCA)、独立成分分析、深度学习和各种形式的聚类算法等。

文献[13]提出一种字典学习和稀疏特征结合的入侵检测模型,该模型包含数据预处理、稀疏特征提取、入侵分类检测和结果分析评估的完整数据处理流程。文献[14]提出深度图特征学习框架 DeepGFL,在网络安全的上下文中提取高阶特征,从低阶网络流特征中导出高阶网络流特征,然后实现网络攻击检测。文献[3]提出一种嵌入二次特征选择的主成分分析特征提取模型。文献[15]通过 PCA 提取表示输入变量变化的相互独立潜在特征,采用基于 MI 特征选择方法选择与模型输出最相关的潜在变量。

3.2 网络流量主要特征提取算法

常用的无监督维数约简技术包括主成分分析、局部线性嵌入(Locally Linear Embedding,LLE)、等度规映射(ISOMAP)等降维算法;监督维数约简技术包括线性判别分析(Linear Discriminant Analysis,LDA)以及近年来比较受关注的度量学习。常用的网络流量特征提取算法及其优缺点和已有研究文献如表 1所示。

表 1 常用网络流量特征提取算法

Table 1 Commonly used network traffic feature extraction algorithms

分类	算法	说明	优缺点	研究文献
监督学习	神经网络	神经网络本身包含一系列特征提取器,理想的 Feature Map 应是稀疏的并且包含典型的局部信息	简单但浅层神经网络提取特征的能力较弱	文献[16-17]
	度量学习	即距离度量学习或相似度学习,可分为通过线性变换的度量学习和非线性模型	可根据不同任务自主学习针对某个特定任务的度量距离函数	文献[18-19]
	线性判别分析	有效利用数据的标签信息,通过最小化同类样本间的差异和最大化不同类样本特征间的差异来提取最佳的判别特征	模型简单,分类效果好,对噪声的鲁棒性较好,但存在特征提取不足和小样本问题,对数据分布做出较强假设,未考虑样本数据的局部结构信息,在实际应用中很受限制	文献[20-21]
无监督学习	局部线性嵌入	通过保留原数据集部分特性的低维数据来重构原始高维数据,从未被标注的高维输入数据中生成低维的近邻保持特征	无需计算距离矩阵,仅需计算稀疏矩阵,可大幅减少计算量,但当数据不满足高斯分布时,性能容易弱化	文献[22-23]
	等度量映射	将 MDS 算法中欧式距离换成测地距离,利用邻接图、Floyd 或 Dijkstra 算法计算出两点之间的最短路径,对离散的样本构造出测地距离	使用“测地距离”而非原始的欧式距离,可更好地控制数据信息的流失,能在低维空间中更全面地表现高维空间数据	文献[18,24]
	主成分分析	使用数据的正交变换,只分析数据的一阶或二阶矩,采用核映射进行拓展得到 KPCA,或采用流形映射对复杂数据集进行非线性降维操作	适用于网络入侵检测,但不适用于分类过程中的特征提取,因为在计算特征轴的最佳旋转时不包含歧视性信息	文献[25-26]
	独立成分分析	使用独立非高斯成分的加权求和技术	存在小样本问题	文献[27-28]
	深度学习	分层结构的神经系统启发多层深度学习架构进行特征学习,输入原始数据,输出低维特征	需要大量的训练数据,且仅适用于包含大量特征的数据集	文献[26,29]

4 网络流量特征选择方法

4.1 网络流量特征选择流程

网络流量特征选择是从原始网络流量特征集中选择出重要的特征,如何选择特征子集以及度量特征的重要性是影响特征选择结果的 2 个重要问题。网络流量特征选择的基本流程如图 5 所示,其中主要包括 4 个环节:生成特征子集,评估特征子集,终止条件判断,验证特征子集。

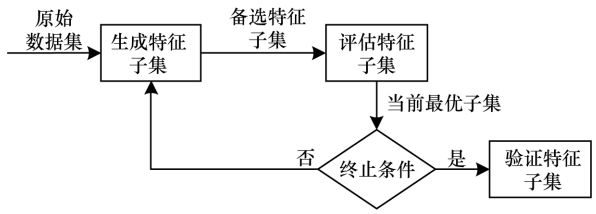

图 5　网络流量特征选择基本流程

Fig. 5　Basic procedure of network traffic feature selection

在图 5 中,原始网络流量数据集需要按照一定的搜索策略生成一个备选网络流量特征子集,根据某个评价准则可判断选出子集的优劣,根据终止条件决定网络流量特征选择算法何时停止,保证算法的有穷性[9]。如果评估结果满足终止条件则停止整个网络流量特征选择过程,否则重复生成网络流量特征子集,循环整个过程。当整个网络流量特征选择过程结束后,需要对所获得的网络流量特征子集进行验证,以证明该网络流量特征选择方法的有效性[30]。

4.2　网络流量特征选择方法分类

网络流量特征选择是指选择最能代表原始网络流量数据分布特性的最优特征子集,根据特征子集评价标准和特征选择形式又可以将网络流量特征选择方法分为 3 类:过滤式(Filter)特征选择方法,封装式(Wrapper)特征选择方法和嵌入式(Embedded)特征选择方法[31]。

1)过滤式特征选择方法。按照发散性或相关性对各网络流量特征进行评分,设定阈值,选择特征。先对网络流量数据进行特征选择,然后再训练学习模型,特征选择过程与后续学习模型无关。该方法优点是速度快、通用性强,而且对过拟合问题有较高的鲁棒性,缺点是特征评价结果有一定的偏差,且模型的组合特征选择能力较差。

2)封装式特征选择方法。从网络流量初始特征集中反复选择特征子集,训练学习模型,根据学习模型的性能对选择出的特征子集进行评价,直至选出最优特征子集。该方法优点是直接针对特定学习器进行优化,考虑到特征之间的关联性,可得到较高的分类精度,缺点是计算复杂度高、开销大,并且通用性不强。

3)嵌入式特征选择方法。使用机器学习算法和模型进行训练,得出网络流量各数据特征的权重,根据权重大小顺序选择特征。该方法与 Filter 方法类似,但其通过机器学习算法和模型训练来确定网络流量特征的优劣,而且算法本身作为组成部分嵌入到学习算法中。最典型的嵌入式特征选择算法是决策树算法,包括 ID3、C4.5 和 CART 算法等。

过滤式和封装式网络流量特征选择方法和分类算法可以与各种算法结合使用,网络流量特征选择过程与学习模型训练过程有明显分别,而嵌入式网络流量特征选择是将特征选择与学习模型训练过程融为一体,在学习模型训练过程中自动地进行特征选择。其中,封装式方法直接将学习器性能作为特征子集的评价标准,搜寻特征子集的分类准确性一般会优于过滤式和嵌入式[32]。

搜索最优网络流量特征子集是网络流量特征选择过程中最关键和最具挑战性的环节。基本搜索策略可根据网络流量特征子集的形成过程分为 3 类:全局最优搜索,随机搜索,启发式搜索。全局最优搜索策略是在所有可能空间中寻找最优子集,针对高维数据,算法的时间复杂度非常高;随机搜索策略使用随机重采样,根据迭代更新特征权重选择重要特征训练分类器,利用模拟退火算法可以避免陷入局部最优解的特性提高搜索性能;启发式搜索策略包括前向选择方法、后向选择方法、序列前向浮动搜索算法等。启发式搜索策略在选择速度上高于前两种搜索策略。一个具体的网络流量特征子集搜索算法可能会采用 2 种或多种基本搜索策略,例如遗传算法是一种随机搜索算法,同时也是一种启发式搜索算法。对于不同的搜索策略,网络流量特征选择方法又可被分为穷举法、启发式法、基于信息理论的方法、基于演化计算方法等[32]。

4.3　网络流量特征选择算法

将过滤式网络流量特征选择方法应用于回归问题时,可使用互信息法;应用于分类问题时,可使用卡方检验法、Relief 方法、方差选择法、相关系数法、互信息法等。封装式网络流量特征选择方法包括 LVW 法、递归特征消除法、穷举法、随机法等。嵌入式网络流量特征选择方法包括正则化法、随机森林、决策树等。常用的网络流量特征选择算法及其优缺点和已有研究文献如表 2 所示。

表 2 常用网络流量特征选择算法

Table 2 Commonly used network traffic feature selection algorithms

分类	算法	说明	优缺点	文献
过滤式网络流量特征选择	互信息法	多数通过消除冗余特征、保留相关特征实现特征选择,互信息用来作为特征和类别的测度,如特征属于该类,互信息量最大,也可用来评价定性自变量与定性因变量的相关性	无需对特征和类别的关系性质作假设,能检测出多种变量间关系,但未考虑特征出现的频率,受边缘概率影响,适合作为分类问题的分类变量筛选方法	文献[33-35]
	卡方检验	可进行独立性检验,检验定性自变量对定性因变量的相关性,使用特征与类别间的关联性来进行量化,关联性越强,特征得分越高,越应被保留	未考虑已选特征和待选特征之间的相关性,不能得到最优解,但速度快,适合离散型特征的选择	文献[36-37]
	信息增益	衡量特征重要性的准则是特征为分类系统带来的信息量,信息越多,该特征越重要	考虑全面、效果好,但仅能衡量特征对整个系统的重要性,而不能具体到某个类别上。适合全局特征选择,无法应用于本地特征选择	文献[16,38]
	Relief/Relief-F 方法	设计一个相关统计量来决定各特征的重要性,通过解决凸优化问题来估计特征权重,特征子集的权重由该子集中各特征所对应的相关统计量分量之和决定	算法有效,成本低,高度耐噪声,且对功能交互有抵抗力,但不稳定,Relief 适合处理二分类问题,Relief-F 可处理多分类问题	文献[39-40]
	相关性特征选择算法	是一种较常用的利用相关性度量的特征选择方法,使用启发式搜索搜索特征子集,并利用相关性对特征子集进行打分,选出较好的特征子集	特征须是离散的随机变量,如是数值型变量,需先执行离散化方法来进行离散化特征	文献[30,41]
封装式网络流量特征选择	遗传算法	作为一种解决最优化的搜索启发式算法,随着算法迭代次数增加,种群适应度值逐渐收敛于局部最优解,从而找到最优特征	具有良好的全局搜索能力,是一种全局优化算法,具有可扩展性,但算法实现复杂,搜索速度比较慢,算法依赖初始种群的选择	文献[42-43]
	随机法	有多种实现方式如 LV/LVW 算法、贪心算法,LV 使用一致性度量作为评价函数,使用 Las Vegas 算法随机搜索子集空间,能很快达到最优解	在高维数据环境下,一般使用贪心算法,LV、LVW 算法在特征比较多时开销很大	文献[44-45]
	墨鱼算法	基于墨鱼的变色性能发现最佳解决方案	能找到最佳解决方案,速度快但检测率低	文献[46-47]
	粒子群优化算法	由鸟类觅食的集群活动启发而提出的启发式算法,是一个 NP-Hard 问题,其在部分解中寻找最优解,是一个局部最优解	计算简单,可同时应用于工程和科研,但易遭受不完全乐观的影响,导致方向和速度调节不够精确,无法解决优化和分散问题	文献[21,48]
	递归特征消除法	使用一个基模型进行多轮训练,每轮训练后,消除若干权值系数的特征,再基于新特征集进行下一轮训练	如数据集维数过高,需花费大量的时间才能得到特征子集	文献[49-50]
嵌入式网络流量特征选择	正则化	把额外约束或惩罚项加到已有模型上,防止过拟合,并提高泛化能力,L1、L2 正则化称为 Lasso、Ridge	稳定性好,L1、L2 正则化都利于降低过拟合风险	文献[51-52]
	随机森林	使用信息增益率进行特征选择,计算每个特征的相对重要性,辅助特征选择,提供 2 种特征选择方法:mean decrease impurity 和 mean decrease accuracy	具有抵抗过拟合以及准确率高、鲁棒性好、易于使用等优点,由于其随机性特点,特征子集的稳定性与一致性不足	文献[48,53-54]
	决策树	算法包含特征选择、决策树的生成与剪枝过程,常用的特征选择度量指标有信息增益、增益率、基尼指数,分别对应 3 种算法:ID3,C4.5 和 CART	适合为数据属性是连续的而非离散的场景,以及多分类问题和回归问题	文献[55-56]
	GBDT	计算特征在单棵树中重要度的平均值,是 Boosting 算法族的一部分,可将弱学习器提升为强学习器的算法,属于集成学习范畴	具有很好的性能,适合进行回归预测	文献[29,57]
	XGBOOST	是 Gradient Boosting 的一种高效实现,并非单一算法,通过每棵树中分裂计算,节点分裂算法能自动利用特征的稀疏性,用于加速和减小内存消耗	可很好地拟合数据,在大数据分析及工业界中表现出色	文献[36,52]

5　网络流量维数约简数据集与评价指标

由于隐私和知识产权等原因,用于网络流量分析的相关数据集较少,很少有公开可用的数据集,且很少提供标记信息。

5.1　网络流量维数约简常用数据集

由于网络设备、流量配置和网络攻击的多样性,任何网络流量数据集的代表性都会被质疑。因此,找到适的标签数据集是很困难的。许多已发表的网络流量异常检测和网络流量维数约简分析的文章仍在使用 DARPA 98 和 KDD CUP 99。常用来研究网络流量维数约简算法使用的网络流量数据集,以及针对该数据集的维数约简方法和已有研究文献如表 3 所示。

表 3　网络流量主要维数约简算法

Table 3　Major network traffic dimensionality reduction algorithms

数据集	年份	说明	维数约简算法
DARPA 98/99	1998 年 1999 年	包括覆盖 Probe、DoS、R2L、U2R 和 Data 5 大类 58 种典型攻击方式,给出 5 周的模拟数据,其中前 2 周是训练数据,后 2 周的数据则用于评测	CFS[41]、BN[58]
KDDcup 99	1999 年	来自 7 周的网络流量的 500 万个连接记录集合,被认为是 DRAPA98 的派生,把各 TCP 数据包集成到 TCP 连接中,含有 39 种不同的攻击手段、4 类异常、不同的网络流量和不同的用户类型	PCA[25]、MI[34]、IG、C4.5 和 BN[33]、CART、BNMB[55]、RF[53]
Kyoto 2006	2006 年	包含由京都大学部署服务器,2006 年 11 月至 2009 年 8 月收集的连续收据,每个连接都有24 个不同特征	MI[35]
NSL-KDD	2009 年	被视为 KDDCUP 99 修订版,解决了测试数据中大量冗余记录和重复记录问题,降低了难度,并提供了更有挑战的攻击分布,包含 41 个基本特性、流量特性和内容特性	IG[38]、MI[35]、GR[16]、DT[56]、GA[42]、ISOMAP[18]
UNB ISCX	2012 年	包含 7 d 的网络活动数据,可用于研究入侵检测和异常检测通用的开发、测试及评估算法	PCA[25]
CAIDA	2014 年	针对特定事件或攻击的许多不同类型的匿名网络跟踪数据,没有标记,缺少多重攻击情形	Chi-Square[37]
CICIDS 2017	2017 年	周一至周五共 5 d 收集的数据,涵盖了代表常见攻击家族的各种攻击情形,包括蛮力攻击、HeartBleed 攻击、僵尸网络、DoS 攻击、DDoS 攻击、Web 攻击和渗透攻击等	PCA、Auto-Encoder[26]、RF[54]、GA[43]

5.2　网络流量维数约简性能评价指标

通常采用分类器准确率(Overall Accuracy,OA)、特征压缩率(Feature Compression Rate,FCR)以及运行时间作为网络流量维数约简算法性能的评价指标。采用分类器准确率评判网络流量维数约简算法效果的好坏,其值为正确样本数与全部样本数的比值。用特征压缩率衡量网络流量维数约简算法对特征提取的效率,其值为选择的特征数与全部特征数的比值。运行时间为每种网络流量维数约简方法所运行的时间,使用每种算法的运行时间来考察其运行速度。

6　网络流量维数约简存在问题及发展趋势

6.1　存在问题分析

当前网络流量异常检测中的维数约简技术已有相关研究,并取得了一定的研究成果,但仍然存在一些尚未解决和完善的问题:传统的维数约简方法无法保留训练样本的原始意义,且对组合特征选择能力较差;网络流量多样性和网络流量数据的不平衡问题,以及复合攻击的普及对网络流量维数约简提出的更高要求;网络加密流量的快速增长需要研究如何从高速网络流量中提取反映加密流量内在规律的特征信息对应的特征提取方法;目前缺乏维数约简评价标准;现有网络流量数据维数约简方法不能正确反映移动无线网络的性能;网络流量的高动态性使得网络流量数据维数约简方法不能满足网络攻击检测的在线实时性要求。上述不足都制约了网络流量异常检测中维数约简技术的进一步发展。

6.2　研究方向展望

基于现阶段网络流量异常检测中维数约简技术的研究现状、网络流量维数约简所面临的挑战和未来研究方向主要概括以下方面:

1)在线实时网络异常检测中流量维数约简技术研究。网络特征建立在海量高速网络流量数据上面,为实现实时在线网络异常检测,需要研究提高网络流量在线时效性的维数约简方法。如何将实时多变量维数约简方法应用到大规模网络流量数据中并对数据进行高效处理成为一大难题。

2)维数约简后流量特征信息丢失问题研究。约简后的网络流量数据特征只是全部网络流量数据特

征的一小部分,一些信息会被丢失。在网络流量异常检测中,如何选择维数约简技术弥补网络流量特征在约简后的信息丢失,使其能有效地进行网络流量异常检测仍是难点。

3)移动互联网应用异常检测中的网络流量特征提取技术研究。随着移动互联网的普及和网络技术的高速发展,移动新应用不断出现,攻击者更青睐于移动互联网应用。如何提取网络流量特征,细分和区别这些网络应用,对攻击检测非常重要。

4)网络流量维数约简评价标准研究。针对网络流量进行有效降维后的特征子集难以确定,缺乏可用于网络流量维数约简的通用和普适的评价标准。

5)网络加密流量的特征提取技术研究。目前缺乏可用于网络加密流量异常检测的公开标记数据集,越来越多的网络流量使用加密通信伪装或隐藏明文流量特征,如何选择待提取的候选特征集合,需对恶意软件加密通信具有全面的知识积累。

6)各种网络攻击检测场景中网络流量数据维数约简技术与方法的普适性问题。目前很多网络流量数据维数约简方法针对某个网络攻击检测场景的应用是最优的,但是针对其他网络攻击检测场景的应用可能就不是最优的。随着针对网络流量数据特征的研究不断深入,未来需要设计普适的网络流量维数约简方法。

7)多种网络流量维数约简方法和技术相结合的维数约简方法。将多种网络流量数据维数约简方法和技术相结合,实现更高效的网络流量数据维数约简和获得更准确的抽样结果。在进行网络流量维数约简时,尽可能地减少对网络的额外影响也是一个具有挑战的研究课题。

7 结束语

网络流量维数约简能够用于很多基于网络流量的机器学习和数据挖掘场景,是网络攻击检测中的重要分支。本文介绍网络流量异常检测和维数约简原理,分别对 2 种流量维数约简方式,即网络流量特征选择和网络流量特征提取的现有算法进行归纳分类,描述算法特点并分析优缺点。在此基础上,给出目前网络流量维数约简研究常用的数据集和评价指标,展望网络流量异常检测中维数约简技术发展方向,为研究和发展网络空间安全技术提供借鉴。

参考文献

[1] JIN Hai. Report on the development of China's cyberspace security cutting edge science and technology in 2018[M]. Beijing: Posts and Telecom Press, 2019. (in Chinese)
金海. 2018 年中国网络空间安全前沿科技发展报告[M]. 北京:人民邮电出版社,2019.

[2] CHEN Liangchen, GAO Shu, LIU Baoxu, et al. Research status and development trends on network encrypted traffic identification[J]. Netinfo Security, 2019, 19(3): 19-25. (in Chinese)
陈良臣,高曙,刘宝旭,等. 网络加密流量识别研究进展及发展趋势[J]. 信息网络安全,2019,19(3):19-25.

[3] CAO Jie. Research of feature reduction and traffic classification method based on SVM[D]. Changchun: Jilin University, 2017. (in Chinese)
曹杰. 基于 SVM 的网络流量特征降维与分类方法研究[D]. 长春:吉林大学,2017.

[4] NI X, HE D, CHAN S, et al. Network anomaly detection using unsupervised feature selection and density peak clustering[C]//Proceedings of International Conference on Applied Cryptography and Network Security. Berlin, Germany: Springer, 2016: 212-227.

[5] NIU Weina. Research on exfiltration complex network attack modeling and identification method[D]. Chengdu: University of Electronic Science and Technology in China, 2018. (in Chinese)
牛伟纳. 窃密型复杂网络攻击建模与识别方法研究[D]. 成都:电子科技大学,2018.

[6] CHEN Liangchen, LIU Baoxu, GAO Shu. Research on traffic data sampling technology in network attack detection[J]. Netinfo Security, 2019, 19(8): 22-28. (in Chinese)
陈良臣,刘宝旭,高曙. 网络攻击检测中流量数据抽样技术研究[J]. 信息网络安全,2019,19(8):22-28.

[7] LUO Ling. Anomaly detection of backbone network based on dimensionality reduction[D]. Hefei: University of Science and Technology of China, 2015. (in Chinese)
罗玲. 基于降维的骨干网流量异常检测研究[D]. 合肥:中国科学技术大学,2015.

[8] WEI Zekun, XIA Jingbo, ZHANG Xiaoyan, et al. Research on traffic multi-feature extraction and classification based on random forest [J]. Transducer and Microsystem Technologies, 2016, 35(12): 55-59. (in Chinese)
韦泽鲲,夏靖波,张晓燕,等. 基于随机森林的流量多特征提取与分类研究[J]. 传感器与微系统,2016,35(12):55-59.

[9] ZHOU Ya. Research on the network traffic identification technology based on semi-supervised learning [D]. Nanjing: Southeast University, 2017. (in Chinese)
周雅. 基于半监督学习的网络业务流量识别方法研究[D]. 南京:东南大学,2017.

[10] HUANG Yinxiang. Research on feature engineering in Internet traffic classification[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2017. (in Chinese)

黄引翔. 网络流量分类中特征工程的研究[D]. 南京:南京邮电大学,2017.

[11] ANUSHA K, SATHIYAMOORTHY E. Comparative study for feature selection algorithms in intrusion detection system[J]. Automatic Control and Computer Sciences,2016,50(1):1-9.

[12] KWON D, KIM H, KIM J, et al. A survey of deep learning-based network anomaly detection[J]. Cluster Computing,2019,22:949-961.

[13] YIN Xiu. Network intrusion detection based on improved dictionary learning[D]. Nanjing: Nanjing University of Posts and Telecommunications,2018. (in Chinese)
尹秀. 基于改进的字典学习的网络入侵检测方法研究[D]. 南京:南京邮电大学,2018.

[14] YAO Yepeng, SU Liya, LU Zhigang. DeepGFL: deep feature learning via graph for attack detection on flow-based network traffic[C]//Proceedings of IEEE Military Communications Conference. Washington D. C., USA: IEEE Press,2018:29-31.

[15] TANG Jian, SUN Chunlai, MAO Kefeng, et al. Network intrusion anomaly detection model based on dimension reduction strategy using principal component analysis and mutual information[J]. Netinfo Security, 2015, 15(9):78-83. (in Chinese)
汤健,孙春来,毛克峰,等. 基于主元分析和互信息维数约简策略的网络入侵异常检测[J]. 信息网络安全,2015,15(9):78-83.

[16] SHRIVAS A K, DEWANGAN A K. An ensemble model for classification of attacks with feature selection based on KDD99 and NSL-KDD dataset[J]. International Journal of Computer Applications,2014,99:8-13.

[17] YE Kai. Key feature recognition algorithm of network intrusion signal based on neural network and support vector machine[J]. Symmetry,2019,11(3):380.

[18] PALECHOR F M, HOZMANOTAS A, HOZFRANCO E, et al. Feature selection, learning metrics and dimension reduction in training and classification processes in intrusion detection systems[J]. Journal of Theoretical and Applied Information Technology,2015,82(2):291-298.

[19] GUPTA J, SINGH J. Detecting anomaly based network intrusion using feature extraction and classification techniques[J]. International Journal of Advanced Research in Computer Science,2017,8(5):1-5.

[20] ABUROMMAN A A, REAZ M B I. Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection[C]//Proceedings of IMCEC'16. Washington D. C., USA: IEEE Press,2016:636-640.

[21] LIU Zhan, JIE Ling, PENG Lin. An industrial control intrusion detection method combining semi-supervised LDA and PSO-SVM[C]//Proceedings of CNCI'19. [S. l.]: Atlantis Press,2019:1-6.

[22] NATESAN P, RAJALAXMI R R, GOWRISON G, et al. Hadoop based parallel binary bat algorithm for network intrusion detection[J]. International Journal of Parallel Programming,2017,45(5):1194-1213.

[23] TANG Jian, ZHUO Liu, JIA Meiying, et al. Supervised nonlinear latent feature extraction and regularized random weights neural network modeling for intrusion detection system[C]//Proceedings of International Conference on Cloud Computing and Security. Berlin, Germany: Springer,2016:343-354.

[24] SUGANYA S, MUTHUMARI G, BALASUBRAMANIAN C. Propitiating behavioral variability for mouse dynamics using dimensionality reduction based approach[C]//Proceedings of ICCTIDE'16. Washington D. C., USA: IEEE Press,2016:1-6.

[25] KEERTHIVASAN K, SURENDIRAN B. Dimensionality reduction using principal component analysis for network intrusion detection[J]. Perspectives in Science,2016(8):510-512.

[26] ABDULHAMMED R, MUSAFER H. Features dimensionality reduction approaches for machine learning based network intrusion detection[J]. Electronics,2019,8(3):322-232.

[27] TSANG C H, KWONG S. Multi-agent intrusion detection system in industrial network using ant colony clustering approach and unsupervised feature extraction[C]// Proceedings of IEEE International Conference on Industrial Technology. Washington D. C., USA: IEEE Press, 2005:51-56.

[28] PALMIERI F, FIORE U, CASTIGLIONE A. A distributed approach to network anomaly detection based on independent component analysis[J]. Concurrency and Computation: Practice and Experience,2015,26(5):1113-1129.

[29] YANG Qing, JIA Cangzhi, LI Taoying. Prediction of aptamer-protein interacting pairs based on sparse autoencoder feature extraction and an ensemble classifier[J]. Mathematical Biosciences,2019(311):103-108.

[30] YU Lei, LIU Huan. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]//Proceedings of ICML'03. Washington D. C., USA: [s. n.], 2003:856-863.

[31] SUN Xingbin, RUI Yun. Feature selection method based on statistic frequency in network traffic classification[J]. Journal of Chinese Computer Systems,2016,37(11):2483-2487. (in Chinese)
孙兴斌,芮赟. 一种基于统计频率的网络流量特征选择方法[J]. 小型微型计算机系统,2016,37(11):2483-2487.

[32] LI Zhanshan, LIU Zhaogeng. Feature selection algorithm based on XGBoost[J]. Journal on Communications, 2019,40(10):101-108. (in Chinese)
李占山,刘兆赓. 基于 XGBoost 的特征选择算法[J]. 通信学报,2019,40(10):101-108.

[33] WANG Wei, HE Yongzhong, LIU Jiqiang, et al. Constructing important features from massive network traffic for lightweight intrusion detection[J]. Constructing IET Information Security,2015,9(6):374-379.

[34] SELVAKUMAR B, MUNEESWARAN K. Firefly algorithm based feature selection for network intrusion detection[J]. Computers and Security,2019,81(3):148-155.

[35] AMBUSAIDI M A,HE X J,NANDA P,et al. Building an intrusion detection system using a filter-based feature selection algorithm[J]. IEEE Transactions on Computers, 2016,65(10):2986-2998.

[36] WANG Jie,XU Jing,ZHAO Cheng'an,et al. An ensemble feature selection method for high-dimensional data based on sort aggregation[J]. Systems Science and Control Engineering,2019,7(2):32-39.

[37] BALKANLI E,ZINCIR-HEYWOOD A N,HEYWOOD M I. Feature selection for robust backscatter DDoS detection[C]// Proceedings of the 40th Local Computer Networks Conference. Washington D.C.,USA:IEEE Press,2015:611-618.

[38] SHADI A, MONTHER A. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model[J]. Journal of Computational Science,2018,25,152-160.

[39] OSANAIYE O, CHOO K K R, DLODLO M. Analysing feature selection and classification techniques for DDoS detection in cloud[C]//Proceedings of Southern Africa Telecommunication Networks and Applications Conference. Fancourt,South Africa:[s.n.],2016:198-203.

[40] HUANG T,CHEN W,ZHANG R. A combined feature selection method based on clustering in intrusion detection[C]//Proceedings of the 2nd International Conference on Automation, Mechanical Control and Computational Engineering. [S.l.]: Atlantis Press, 2017:2352-5401.

[41] KHOR K C,TING C Y,AMNUAISUK S P. A feature selection approach for network intrusion detection[C]// Proceedings of the 11th International Conference on Information Management and Engineering. London,UK: [s.n.],2019:133-140.

[42] BENAICHA S E,SAOUDI L,GUERMECHE S E B,et al. Intrusion detection system using genetic algorithm[C]// Proceedings of Science and Information Conference. London, UK:[s.n.],2014:564-568.

[43] VIJAYAN R, DEVARAJ D, KANNAPIRAN B. Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection[J]. Computers and Security,2018,77,304-314.

[44] EL-ALFY E S M,AL-OBEIDAT F N. A multicriterion fuzzy classification method with greedy attribute selection for anomaly-based intrusion detection[J]. Procedia Computer Science,2014,34:55-62.

[45] BAHL S,SHARMA S K. A minimal subset of features using correlation feature selection model for intrusion detection system[C]//Proceedings of the 2nd International Conference on Computer and Communication Technologies. Berlin,Germany:Springer,2016:337-346.

[46] EESA A S,ORMAN Z,BRIFCANI A M A. A novel feature selection approach based on the cuttlefish optimization algorithm for intrusion detection systems[J]. Expert System with Applications,2015,42(5):2670-2679.

[47] KAMBATTAN K R,MANIMEGALAI R,GANAPATHY S. An increment feature selection approach for intrusion detection system in MANET[J]. International Journal for Research in Applied Science and Engineering Technology, 2017,5(1):325-329.

[48] MALIK A J, SHAHZAD W, KHAN F A. Network intrusion detection using hybrid binary PSO and random forests algorithm[J]. Security and Communication Networks,2015,8(16):2646-2660.

[49] USTEBAY S, TURGUT Z, AYDIN M A. Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier[C]// Proceedings of International Congress on Big Data,Deep Learning and Fighting Cyber Terrorism. Washington D.C., USA:IEEE Press,2018:71-76.

[50] SHAH R,QIAN Y,KUMAR D,et al. Network intrusion detection through discriminative feature selection by using sparse logistic regression[J]. Future Internet, 2017,9(4):81.

[51] GHOSH P, MITRA R. Proposed GA-BFSS and logistic regression based intrusion detection system[C]//Proceedings of the 3rd International Conference on Computer, Communication, Control and Information Technology. Washington D.C.,USA:IEEE Press,2015:1-6.

[52] ZHENG Huiting,YUAN Jiabin,CHEN Long. Short-term load forecasting using EMD-LSTM neural networks with an Xgboost algorithm for feature importance evaluation[J]. Energies,2017,10(8):1168-1176.

[53] HASAN M,NASSER M. Feature selection for intrusion detection using random forest[J]. Journal of Information Security,2016(7):129-140.

[54] SHARAFALDIN I,LASHKARI A H,GHORBAN A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. Funchal,Portugal:[s.n.],2018: 22-24.

[55] ALDWAIRI M, KHAMAYSEH Y, AL-MASRI M. Application of artificial beecolony for intrusion detection systems[J]. Security and Communication Networks, 2015,8(16):2730-2740.

[56] POPOOLA E,ADEWUMI A. Efficient feature selection technique for network intrusion detection system using discrete differential evolution and decision tree[J]. International Journal of Network Security,2017,19(5): 660-669.

[57] RAO H D,SHI X Z,RODRIGU A K. Feature selection based on artificial bee colony and gradient boosting decision tree[J]. Applied Soft Computing,2019(74): 634-642.

[58] CHEBROLU S,ABRAHAM A,THOMAS J P. Feature deduction and ensemble design of intrusion detection systems[J]. Computers and Security,2005,24(4): 295-307.

编辑 金胡考